Assignment-based Subjective Questions:

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

There are seven numbers of categorical features present in the dataset. They are Season, Year, Month, Holiday, Weekday, Workingday, and Weathersit. The analysis of the features has been done based on Box Plot, Bar Plot. The points based on which we can infer about their effect on the dependent variable are mentioned below:

- **Season** - count of bike sharing is maximum for fall which is followed by Summer & Winter and least for spring. The booking count has been increased every year for all the season
- **Year** - count of bike sharing increased in 2019 which shows good progress in business.
- **Month** - count of bike sharing increased in the month of May, June, July, August, September, October with a median of over 4000 bookings per month
- **Holiday** - Maximum number of bike rentals are happening during non-holiday time
- **Weekday** - There is no much variance in number of bike sharing throughout the Weekdays as it does not show any specific trend here
- **Workingday** - count of bike sharing is high for weekdays
- **Weathersit-** Clear/Few clouds/Partly cloudy/Partly cloudy weathers show a positive trend in the number of bike users

2. Why is it important to use drop_first=True during dummy variable creation?

drop_first = True is important to use when we change a categorical variable into dummy variables because the last category is already indicated by having a 0 on all other dummy variables. Including the last category just adds redundant information, resulting in multicollinearity. Apart from that, it also reduces the extra column created during dummy variable creation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

[‘registered’, 'casual','temp','atemp']-These are the variables which are highly correlated with the target variable (‘cnt’).

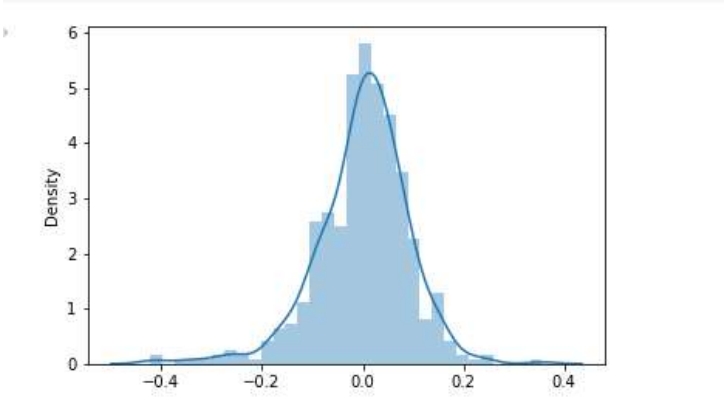
Casual & registered both the columns contains the count of bike booked by different categories of customers. From the above analysis we can conclude that total bike rental value 'cnt = 'casual' + 'registered'. Since we can ignore these two columns. So we can conclude “temp” and “atemp” are highly correlated with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Linearity:
  - We use a pair plot to check the relation of independent variables with the dependent variable.

- Homoscedasticity**:**
  - Homoscedasticity means that the residuals have equal or almost equal variance across the regression line. By plotting the error terms with predicted terms we can see that there is not any pattern in the error terms

- Normality of residuals:
  - The residuals terms are normally distributed and have a mean value of zero.



- Multicollinearity:
  - All the features have VIF value less than 5. So we can consider that there is insignificant multicolinearity among the features.

- No autocorrelation of residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top 3 features of the final model are:

- temp : A coefficient value of 0.479893

- yr : A coefficient value of 0.234324

- Light Snow/Light Rain + Thunderstorm + Scattered clouds/Light Rain + Scattered clouds : A coefficient value of -0.286518

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range.

Mathematical equation of linear regression:

$Y = mX + b$

Y = the dependent variable we are trying to predict.
X = the independent variable we are using to make predictions.
m = the slope of the regression line which represents the effect X has on Y
b = the constant, known as the intercept. If X = 0, Y would be equal to b.

Linear regression are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. X represents our input data and Y represents our prediction.

$$Y = mX + b$$

Multivariable regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, the model will try to learn.

$$Y = w0+w1x1+w2x2+w3x3+\ldots$$

w0 = the intercept

w1 = coefficient for x1 variable

w2 = coefficient for x2 variable and so on.

The variables $x1, x2, x3$ represent the attributes, or distinct pieces of information.

Cost function:

To achieve the best-fit regression line it is very important to update the weights (w1,w2,..) values, so that the error difference between predicted value and true value is minimum. MSE measures the average squared difference between an observation's actual and predicted values.

$$MSE_{minimize} = {}^{1}/_{n} \sum_{1}^{n} (y_i - (mx_i + b))^2$$

The Assumptions of Linear Regression:

1. Linear relationship**:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2. Independence**:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3. Homoscedasticity**:** The residuals have constant variance at every level of x.

4. Normality**:** The residuals of the model are normally distributed.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

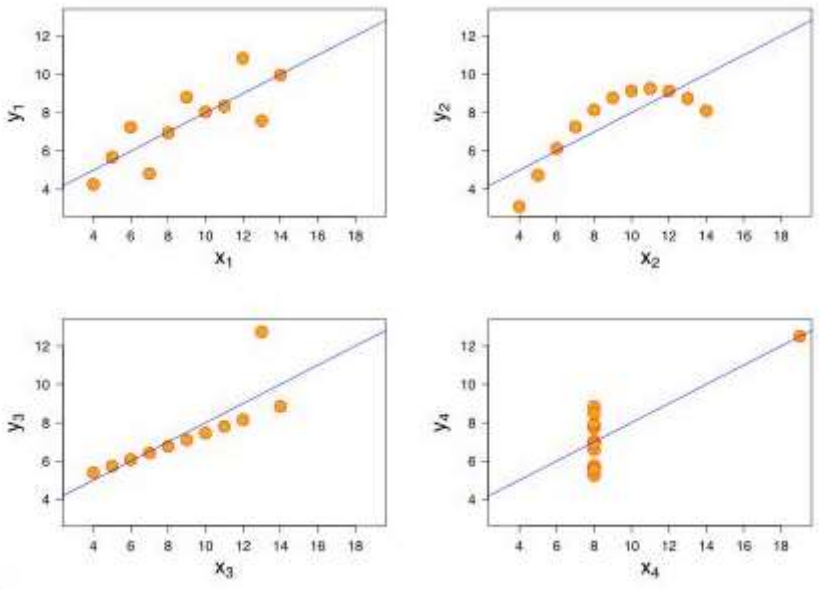
The Datasets are:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The Statistical summary of the datasets:

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

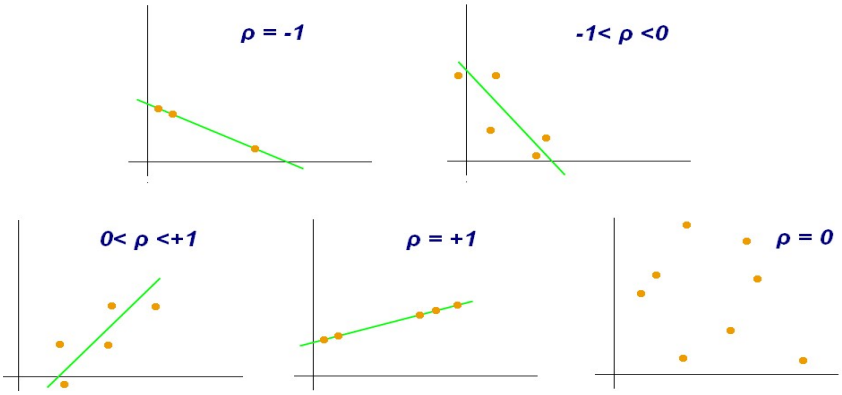
Mean of X is 9 and mean of Y is 7.50 for each dataset. Similarly, the std of X is 3.32 and std of Y is 2.03 for each dataset .The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.



- In the Dataset I, there seems to be a linear relationship between x and y.
- In the Dataset II, there is a non-linear relationship between x and y.
- In the Dataset III, there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- In the Dataset IV shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

The Pearson's Correlation Coefficient is referred to as **Pearson's r,** it is a statistic that measures the linear correlation between two variables and it has a numerical value that lies between -1.0 and +1.0.



r/p = 1 means the data is perfectly linear with a positive slope
r/p = -1 means the data is perfectly linear with a negative slope
r/p = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem. These algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

- Difference between normalized scaling and standardized scaling

Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them.

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardization:

This is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. It translates the data to the mean vector of original data to the origin and squishes or expands the points if standard deviation is 1 respectively.

```
X_new = (X - mean)/Std
```

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset.

$$VIF_i = \frac{1}{(1-R_i^2)}$$

VIF_i is the value of VIF for the i th variable, R_i^2 is the R2 value of the model when that variable is regressed against all the other independent variables .Now if VIF_i is infinite then $(1 - R_i^2)$ is zero, then we can conclude R_i^2=1 which shows a perfect correlation between independent variables means ,these independent variable can be explained perfectly by other independent variables. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. One of the assumptions of linear regression is the residuals of the model are normally distributed. To check the distribution of the residuals, we can use quantile-quantile (q-q) plot.